36-402 DA Exam 1

Keltin Grimes (kgrimes)

March 25, 2022

Introduction

Our client is a large technology company who is considering moving their headquarters to a large city. The company would like to present a case to the city government that moving their headquarters there is deserving of a large tax-break. To do this we need to examine what factors affect the economic output of cities. We measure economic output in terms of per-capita Gross Metropolitan Product (GMP). We have two candidate hypotheses for how per-capita GMP is related to population: the power-law-scaling hypothesis which suggests that per-capita GMP increases exponentially with population; and the urban hierarchy hypothesis which suggests that GMP is not related to population, only the number and size of economically productive companies in the city. Moving the company's headquarters to a large city would not have much effect on the overall population of the city, so if the power-law-scaling model holds, the new headquarters would be of little benefit to the city. If the urban hierarchy model is correct, then the addition of this highly productive company should increase the city's GMP. If we can show that the new headquarters would actually increase GMP, then the company would have a strong argument for receiving tax breaks from the city. **(1)** Therefore in this report we have aimed to evaluate the validity of the power-law-scaling model both before and after accounting for various economic variables beyond population, and discuss what the results mean for our client.

(2) We analyzed two models of per-capita GMP, the power-law-scaling model, and an additional model based on four variables dealing with various economic sector data. The power-law-scaling model was no better than this additional model. Additionally, we found that, after accounting for the four economic sector variables, the power-law-scaling hypothesis does not actually hold. Our evidence suggests that per-capita GMP actually decreases with larger populations after adjusting for the variables, but this effect is minimal for the types of large cities our client is considering moving to. We can say with confidence

that the power-law-scaling model is not valid, and that the urban hierarchy hypothesis is likely correct.

Exploratory Data Analysis

We have a dataset consisting of 133 Metropolitan Statistical Areas (MSAs), each with its per-capita Gross Metropolitan Product (GMP) in US dollars, population, and the proportion of the economy in each of the following sectors: finance; professional and technical services; information, communication, and technology (ICT); and corporate management. Population will be our key variable of interest as we attempt to assess the validity of the power-law-scaling and urban hierarchy hypotheses. **(1)** From Figure 1, we can see that population is unimodal, centered around a median of 183,300, but with a very long tail of MSAs with very large populations. The four economic sector variables are also unimodal with varying degrees of rightward skew. We present in Table 1 the mean and median of all the predictor variables.

Table 1: Observed Centers of Predictor Variables.

	Population	Finance	Prof. and Technical	ICT	Management
Mean	450486	0.149	0.047	0.040	0.009
Median	183300	0.135	0.041	0.021	0.007

(2) Our goal is to model GMP, and as we mentioned we have the per-capita GMP in US dollars for each of the MSAs. From Figure 1 we can see that the distribution of per-capita GMP is unimodal and centered around $32,000, with skew towards higher values. There also appears to be one outlier MSA with a per-capita GMP of about $77,000, over $20,000 more than the next highest city.

As we will discuss in more detail in the Modeling & Diagnostics section, we will evaluate the power-law-scaling hypothesis by constructing a linear model of log population versus log per-capita GMP. We plot the relationship between the two transformed variables in Figure 2. **(3)** As we can see, after the transformations the relationship looks fairly linear, which is what we would expect if the power-law-scaling model was valid. We would also like to adjust for the other variables we have, to determine if the power-law-scaling model holds even when accounting for potential confounders. We display scatter plots of per-capita GMP and the four economic sector variables in Figure 3. We can see that the

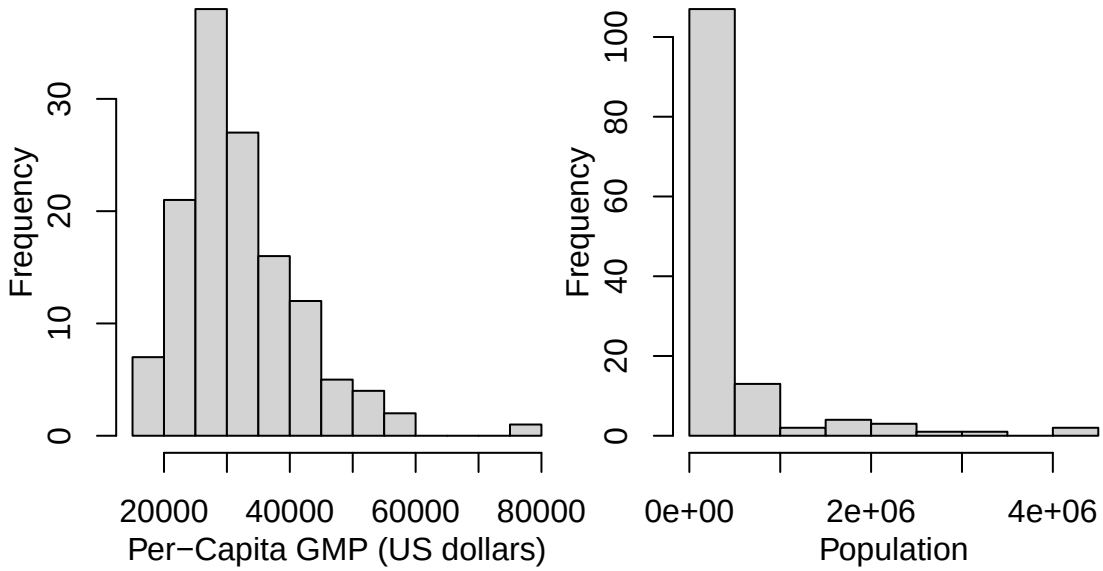

Figure 1: Histograms of Per-Capita GMP (left) and Population (right).

relationships between per-capita GMP and the proportion of the economy in Finance, as well as Professional and Technical Services, are fairly linear. However, for per-capita GMP and both ICT and Corporate Management, the relationship appears to be non-linear. We may have to consider transforming these two variables when constructing our models, or using a non-linear model. **(4)** We also examined the relationship between the predictor variables, and found that the correlations are weak enough that we should not have to worry about dealing with interaction terms.

Modeling & Diagnostics

To properly model the power-law-scaling hypothesis we need a more formal definition of its claim. Let Y represent per-capita GMP, and let N represent a city's population. Then the power-law-scaling hypothesis asserts that $Y = bN^a$ for some $a > 0$ and $b > 0$. We want to assess how well this model fits our data, but this exponential form is not ideally suited for statistical modeling. By taking the log of both sides of the equation, we get an equivalent form of $\log(Y) = c + a\log(N)$ for some constant c. **(1)** This form allows us to construct the linear model:

$$\log(\text{per-capita GMP}) = \beta_0 + \beta_1 \log(\text{Population})$$

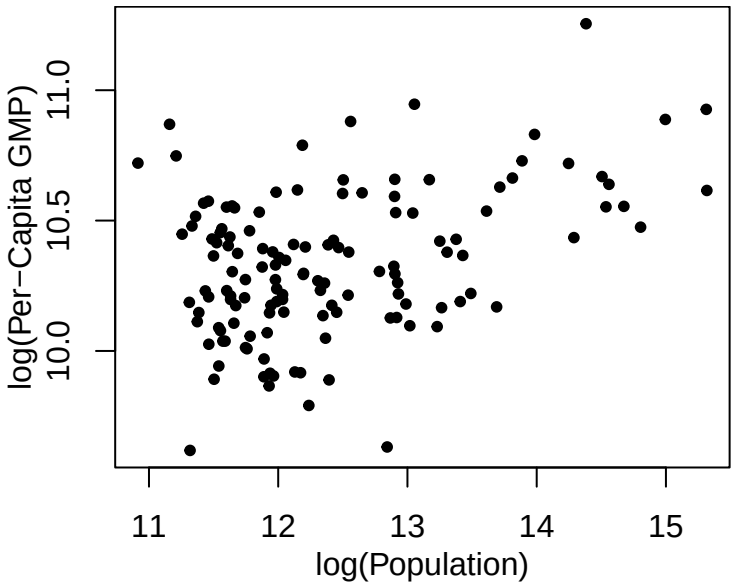

Figure 2: Scatterplot of log(Population) vs. log(per-capita GMP).

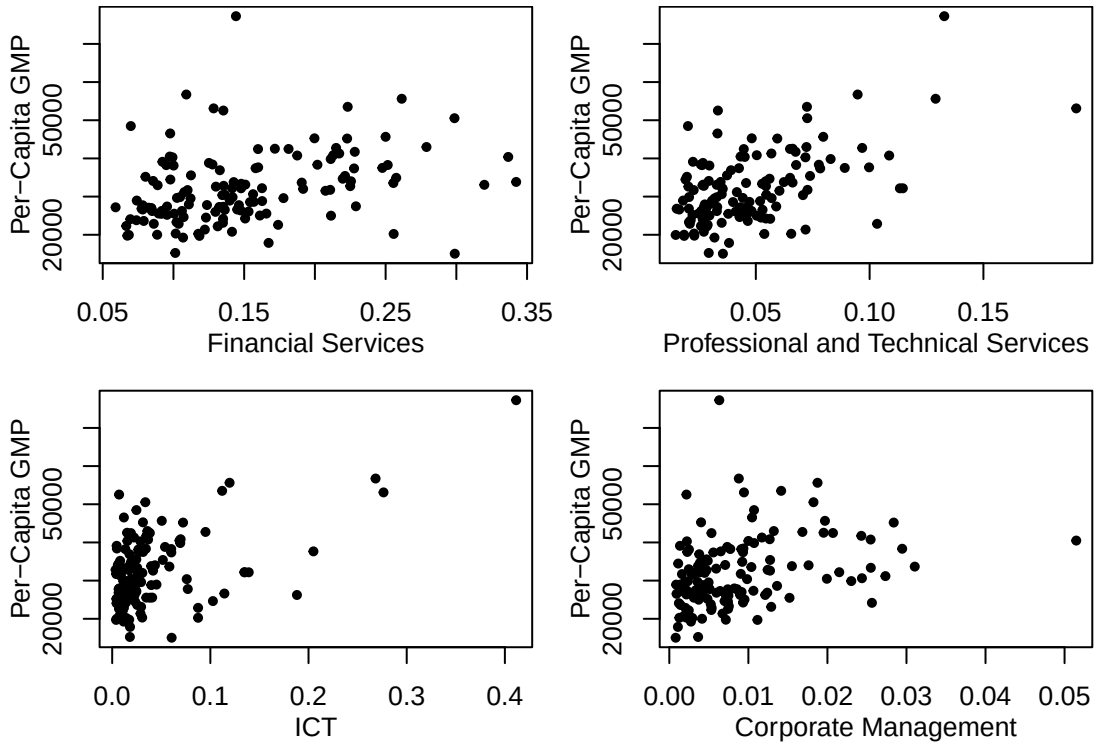

Figure 3: Scatterplots of per-capita GMP and the economic sector variables. Each economic variable is measured in terms of proportion of entire economy.

and using linear regression we estimate $\hat{\beta_0} = 8.903$ and $\hat{\beta_1} = 0.115$. We refer to this as Model 1. **(2)** From Figure 4 we can see that the residuals appear to be normally distributed with mean zero, but the variance is not constant across the fitted values. Specifically, the variance of the residuals tends to decrease as population increases. We did not observe any outliers, despite the inclusion of the MSA with very high per-capita GMP pointed out previously. **(3)** We use 10-fold cross validation to estimate the prediction error of our models (results are shown in Table 2 in the Results section). This is an approximately unbiased estimate of the true prediction risk, but is still a random quantity so carries some uncertainty.

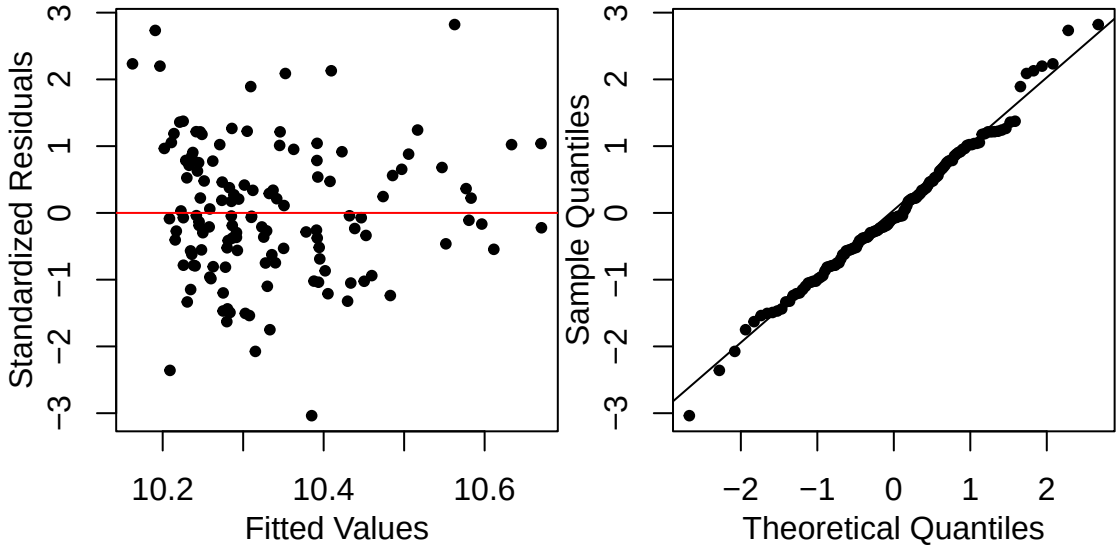

Figure 4: Model 1 (power-law-scaling model) residual plots. Scatterplot of fitted values vs. standardized residuals (left), and QQ-Plot of residuals (right).

Converting the power-law-scaling model into a linear model is convenient, but doing so introduces bias into the model, because exponentiating the predicted values is not a perfect reconstruction of the true per-capita GMP, in expectation. In general, the linear model will tend to underestimate the per-capita GMP. We would like to quantify this bias for cities of the type our client is considering moving to. We choose Pittsburgh as a representative city, which has a population of 2,361,000, so will estimate the bias of Model 1 on cities of that size. We will estimate the bias by repeatedly refitting the model in a bootstrap procedure. **(4)** Since the variance of the residuals is not constant over MSAs with different population, we will bootstrap the model's predictions by resampling the data with replacement. We follow the pivotal assumption, that the difference between our estimate of per-capita GMP and the true mean per-capita GMP is roughly equivalent in distribution to our mean

bootstrapped per-capita GMP and our original estimate of per-capita GMP. So our estimate of the bias will be the mean of the boostrapped predictions, minus our prediction from the original model.

We would also like to build a model of per-capita GMP from the four economic sector variables, so that we can adjust for them in our power-law-scaling model. **(5)** Due to the non-linearities in the relationship between per-capita GMP and the ICT and Corporate Management sector variables, we will fit a non-linear model, specifically a kernel regression model with a Gaussian kernel. We use an ad-hoc method to select the tuning parameters, rather than using an expensive cross-validation procedure. We call this Model 2. We plot the residuals for this model in Figure 5, and we can see that this model also has issues with non-constant variance in the residuals, which are also less normally distributed than those of Model 1.

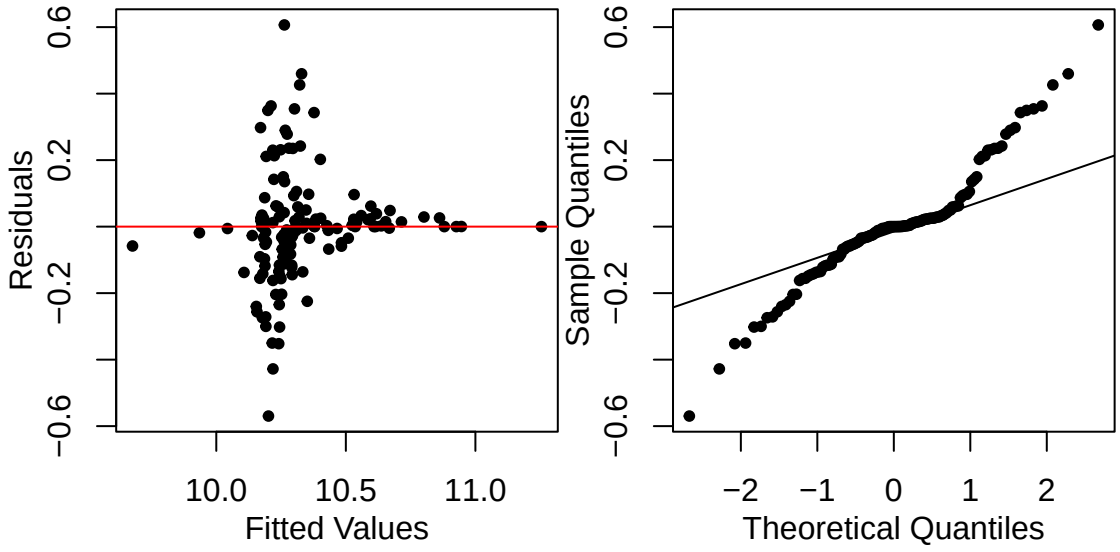

Figure 5: Model 2 (kernel regression model) residual plots. Scatterplot of fitted values vs. standardized residuals (left), and QQ-Plot of residuals (right).

We then fit a linear model on log population to predict the residuals of Model 2, which is equivalent to fitting the power law scaling model while accounting for the four economic sector variables. **(6)** The model is of the form:

$$\text{Model 2 Residuals} = \beta_0' + \beta_1' \log(\text{Population})$$

where Model 2 is the kernel regression model fitted on the economic sector variables. We use least squares to get the parameter estimates $\beta_0' = 0.295$ and $\beta_1' = -0.024$.

Results

Table 2 shows the results of our 10-fold cross validation procedure on the original power-law-scaling model (Model 1). **(1)** We found a Mean-Squared Error of 74898951 with a standard error of 37057351. Although our model is not optimized for Mean Absolute Error, it is useful to examine since it is given in the same units as per-capita GMP (dollars), and should be easier for policy makers to interpret than MSE. The model's cross-validated MAE is \$6486, with a standard error of \$1224. Using the bootstrap procedure described in the Modeling & Diagnostics section, we estimate that the bias of Model 1 in predicting the per-capita GMP of cities with the same population as Pittsburgh to be approximately -\$13. This tells us that, on average, our model will underestimate the true per-capita GMP of cities with a population of 2,361,000 by about thirteen dollars. Relative to the observed values of per-capita GMP in our dataset, whose minimum value is about \$15,000, and based on our background knowledge about the wealth of cities in the United States, this bias is pretty minor. Ultimately, this model is not very useful beyond giving a fairly rough estimate of the per-capita GMP of a city. The error is fairly high, and as discussed in the previous section, the variance of the residuals for this model are non-constant, which violates the assumptions of linear regression.

Table 2: 10-fold Cross Validation estimates and standard errors of MSE and MAE.

	MSE	MSE Std. Error	MAE	MAE Std. Error
Model 1	74898951	37057351	6486	1224
Model 2	59553701	31094151	5699	1293

The results of our cross-validation procedure on Model 2, the kernel regression on the economic variables, are also presented in Table 2. **(2)** We found a Mean-Squared Error of 59553701 with a standard error of 31094151. Again we report the MAE results for interpretability's sake. The model's MAE is \$5699, with a standard error of \$1293. The MSEs of the two models are less than one standard error away from each other, which tells us the error of the two models are not significantly different. The same goes for the MAE, but we are less interested in the statistical significance of this measure of error, and more concerned of its practical significance. The key conclusion to make from this is that the power-law-scaling model is *not any better* than a model only based on four economic sector variables, which gives us evidence for the urban scaling hypothesis.

(3) Based on Model 3, the regression of log population on the residuals of Model 2, our estimate of the scaling exponent in the power-law-scaling model is -0.0239. In other words, after accounting for the four economic sector variables, our estimate of the scaling exponent a is -0.0239. We bootstrap our model fitting procedure (both Model 2 and 3) by resampling the data with replacement 1000 times in order to estimate the uncertainty of this prediction. Doing so, we find a pivotal confidence interval of our estimate of the scaling exponent to be $(-0.0405, -0.0041)$. Our goal of this report was to assess the validity of the model $Y = bN^a$, where $a > 0$ and $b > 0$. **(4)** Clearly it is not the case that $a > 0$ after accounting for the economic sector variables, so we should reject the power-law-scaling model hypothesis. This gives us reason to believe that the urban hierarchy hypothesis is the correct model of the relationship between population and GMP.

We should note that the confidence interval of our estimate of the scaling exponent actually suggests that the model $Y = bN^a$ is valid, but for some $a < 0$. This would imply that per-capita GMP, after accounting for the four economic sector variables, actually decreases as population increases. However our estimated values for a and b are small enough that this effect should be largely negligible when assessing the effect of our client moving to a new city, especially if that city already has a high population, which we know will be the case.

Conclusions

(1) Our report found that, on the surface, the power-law-scaling model appears to be a decent predictor of per-capita GMP, although it is systematically slightly biased due to its formulation. For cities the size of Pittsburgh, this bias results in underestimates of about \$13. The predictive performance of this model would suggest the power-law-scaling model is a reasonably valid model of the relationship between population and GMP. However, after accounting for various other economic variables, we find that the power-law-scaling model clearly does not hold, and that these other variables were confounding the original model. So we can say with confidence that the power-law-scaling hypothesis is not valid. This suggests that the urban hierarchy hypothesis is a much better model of GMP. We did find that, after adjusting for the other economic variables, per-capita GMP *decreased* as population increased, but this effect is only noticeable for cities with very small population. Since our client is considering moving to a large city and will have little impact on the population, this should not be a concern.

(2) Although we can confidently disprove the power-law-scaling hypothesis and have

provided evidence in support of the urban hierarchy hypothesis, we have not made any claims about the exact nature of the urban hierarchy hypothesis, so are unable to provide any concrete claims about how our client's new headquarters will improve the economy and increase the GMP of the chosen city. Neither of the models we built will predict accurately enough to provide a good estimate of per-capita GMP of cities not in the dataset, and will have an even harder time quantifying the change in GMP resulting from our client's new headquarters. We recommend investigating the urban hierarchy hypothesis further, ideally with additional predictor variables, to get a concrete estimate of the new headquarters' value to potential host cities.